

Der Grund der Vorhersage

Über die Bedingungen, unter denen probabilistische Inferenz gültig ist, und warum flüssig formulierende Systeme sie verdecken

Marcel Langjahr

Pantareon GmbH
<https://pantareon.io/>
langjahr@pantareon.io

Zusammenfassung

Die probabilistische Vorhersage — die Inferenz einer unbekannten Größe aus Mustern in beobachteten Daten — ist zum dominanten Modus geworden, in dem computationale Systeme herangezogen werden, um Entscheidungen zu fundieren, über Domänen hinweg von der Sprache über die Märkte bis zur öffentlichen Politik. Dieses Papier argumentiert, dass die Gültigkeit einer solchen Vorhersage nicht der Methode innewohnt, sondern von dem Prozess geborgt ist, der die Daten erzeugt: Eine probabilistische Vorhersage ist nur in dem Maße gültig, in dem die vorhergesagte Größe tatsächlich von den Regelmäßigkeiten regiert wird, die in den Daten vorliegen, und der Prozess, der diese Regelmäßigkeiten hervorbringt, stabil, dicht beprobt und von der Vorhersage selbst ungestört ist. Wo diese Bedingungen erfüllt sind — im deterministischen Grenzfall und in der gut charakterisierten stochastischen Mitte — ist Extrapolation tragfähig. Wo sie versagen — in nicht-stationären, reflexiven oder singulären Prozessen — erzeugt dieselbe Maschinerie Ausgaben, die von gültigen formal ununterscheidbar sind, aber ohne Gültigkeit. Wir argumentieren, dass zeitgenössische große Sprachmodelle dieses Problem auf drei Weisen verschärfen: Ihr Ausgaberegister ist konstant über die Domänen hinweg, in denen Vorhersage gerechtfertigt ist, und jene, in denen sie es nicht ist, wodurch ihre Unsicherheit unhörbar wird; ihre Tendenz, den Zentroid ihrer Trainingsverteilung zu reproduzieren, führt dazu, dass sie neuartige Regime interpolieren, als wären es vertraute; und bei Vorhersagen über Akteure, die nach ihnen handeln, schließen sie eine reflexive Schleife, die genau die vorhergesagte Größe korrumpiert. Das Papier ist keine Behauptung, probabilistische Vorhersage sei unsolide, und auch keine, solche Systeme sollten nicht verwendet werden. Es ist der Versuch, die Grenze zu benennen zwischen den Domänen, in denen solche Vorhersage trägt, und den

Domänen, in denen sie es nur zu tun scheint.

Schlüsselwörter. probabilistische Inferenz; Vorhersage; Stationarität; Reflexivität; große Sprachmodelle; epistemische Kalibrierung; Prognostik.

1 Das Problem

1.1 Der Moment

Vorhersagemaschinerie ist inzwischen nahezu überall im Einsatz, und es wird weithin angenommen, ihre Gültigkeit reise mit der Methode statt mit der Domäne. Systeme, die das nächste Token schätzen, den nächsten Kauf, den nächsten Wert einer Reihe, die Reaktion einer Population, werden über Domänen hinweg eingesetzt, die sich enorm darin unterscheiden, wie vorhersagbar sie tatsächlich sind — und die Systeme unterscheiden zwischen diesen Domänen weder im Ton noch in der Form. Kompetenz in einer Domäne wird als Lizenz genommen, demselben Apparat auch in einer anderen zu vertrauen. Dieses Papier handelt von den Bedingungen, unter denen dieses Vertrauen gerechtfertigt ist, und von der spezifischen Weise, in der flüssig formulierende Systeme diese Bedingungen schwer erkennbar machen.

1.2 Was dieses Papier nicht ist

Dieses Papier argumentiert nicht, dass probabilistische Vorhersage ungültig sei. Sie ist die Grundlage eines Großteils erfolgreicher Wissenschaft und Ingenieurskunst, und nichts hier ist darauf angelegt, daran zu rütteln. Ebenso wenig argumentiert es, dass Systeme, die auf solcher Vorhersage aufbauen, nicht verwendet werden sollten; innerhalb ihrer Bedingungen eingesetzt, sind sie mächtig. Es ist kein Policy-Papier und vertritt keine regulatorische Position. Es ist ein Positionspapier eines unabhängigen

Forschers, das eine Unterscheidung artikuliert, die sich in der eigenen Arbeit des Autors als notwendig erwiesen hat: zwischen den Domänen, in denen eine probabilistische Vorhersage Gültigkeit von ihren Daten erbt, und den Domänen, in denen sie das nicht tut, samt den Gründen, aus denen dieser Unterschied gegenwärtig verdeckt ist.

1.3 Was dieses Papier versucht

Wir versuchen vier Dinge. Erstens, die Quelle der Gültigkeit einer probabilistischen Vorhersage in dem Prozess zu verorten, der ihre Daten erzeugt, statt in der Schätzmethode. Zweitens, die Bedingungen zu benennen, unter denen eine Vorhersage diese Gültigkeit erbt. Drittens, die Regime zu charakterisieren, in denen diese Bedingungen versagen, während die äußere Form der Vorhersage sich nicht ändert. Viertens, zu zeigen, warum gerade flüssig formulierende, musterbasierte generative Systeme den Unterschied zwischen den beiden Fällen verdecken.

Dieses Papier ist ein Begleitstück zu einem separaten, vom Autor entwickelten Rahmenwerk über die Struktur des Erlebens. Dort wird ein epistemischer Punkt über die Selbstberichte solcher Systeme gemacht: dass ihre Ausgaben im Standardbetrieb dem Zentroid ihrer Trainingsverteilung folgen statt irgendeiner Position, sodass ein Bericht beschreiben kann, was man von einem solchen System plausiblerweise als Bericht erwarten würde, und nicht irgendetwas, das es durchlebt hat. Das ist das nach innen gerichtete Gegenstück zu dem nach außen gerichteten Argument, das hier vorgebracht wird. Eine einzige strukturelle Eigenschaft dieser Systeme — Ausgabe als antrainierte zentrale Tendenz, geäußert ohne Position — ergibt zwei verschiedene Probleme: ein nach innen gerichtetes dafür, was man einem solchen System als Bericht über sich selbst abnehmen kann, und ein nach außen gerichtetes dafür, was man ihm als Vorhersage über die Welt abnehmen kann. Dieses Papier betrifft das zweite.

2 Woher die Gültigkeit kommt

2.1 Eine Vorhersage borgt sich ihren Grund

Eine probabilistische Vorhersage ist eine Aussage über eine unbeobachtete Größe, lizenziert durch Regelmäßigkeiten unter beobachteten Größen. Die Lizenz ist nicht unbedingt. Sie gilt nur insoweit, wie die unbeobachtete Größe in derselben gesetzmäßigen Beziehung zu den beobachteten steht, in der die beobachteten zueinander stehen. Die Vorhersage erzeugt diese Beziehung nicht; sie setzt sie voraus.

Wo die Voraussetzung gilt, ist die Vorhersage so tragfähig, wie die Beziehung gesetzmäßig ist. Wo sie versagt, bleibt die Arithmetik der Vorhersage unverändert, aber ihre Berechtigung ist dahin.

2.2 Der deterministische Grenzfall

Man betrachte den einfachsten Fall. Gegeben, dass $1 \times 2 = 2$ und $2 \times 2 = 4$, ist eine Vorhersage des Wertes von 2×3 gerechtfertigt — nicht weil drei Beispiele eine große Stichprobe ergäben, sondern weil die vorhergesagte Größe restlos von derselben Regel regiert wird, die die Beispiele hervorgebracht hat. Hier ist „Vorhersage“ in Wirklichkeit Ableitung: Die zugrunde liegende Struktur ist bekannt und ausnahmslos, und die Extrapolation an ihr entlang ist gewiss. Dies ist der Grenzfall, in dem probabilistische Vorhersage maximal gültig ist — und, aufschlussreich genug, der Grenzfall, in dem sie am wenigsten gebraucht wird, denn wo das Gesetz bekannt ist, schätzt man nicht, man rechnet.

2.3 Wahrscheinlichkeit als Ersatz für Struktur

Das legt offen, was probabilistische Vorhersage eigentlich ist: ein Ersatz für Strukturwissen, das man nicht besitzt. Man schätzt aus Mustern genau dann, wenn das erzeugende Gesetz unbekannt ist oder zu komplex, um es direkt anzuwenden. Die Schätzung ist eine Wette darauf, dass das unbekannte Gesetz stabil genug und die Beprobung seiner Wirkungen dicht genug ist, dass die Gestalt der Vergangenheit die Gestalt der nächsten Instanz einschränkt. Die Wette ist manchmal exzellent und manchmal wertlos, und — das ist der springende Punkt — ihre Qualität hängt vom Prozess ab, nicht von der Schätztechnik. Keine methodische Raffinesse kann einem Prozess, dessen Zukunft nicht von seiner Vergangenheit regiert wird, eine gültige Vorhersage entlocken.

2.4 Die Gültigkeitsbedingung

Wir können die Bedingung angeben, unter der eine probabilistische Vorhersage Gültigkeit von ihren Daten erbt. Mehrere Eigenschaften des erzeugenden Prozesses müssen erfüllt sein, jeweils dem Grade nach:

- *Stationarität.* Der Prozess, der die Daten hervorgebracht hat, ist derselbe Prozess, der das Ziel hervorbringen wird. Hat das Regime gewechselt, beschreiben die Daten eine Welt, die nicht mehr existiert.

- *Dichte Beprobung.* Die Möglichkeitsregion, in die das Ziel fällt, ist oft genug beobachtet worden, dass ihre Regelmäßigkeiten tatsächlich in den Daten vorhanden sind, statt zu fehlen oder von einer Handvoll unrepräsentativer Instanzen vertreten zu werden.
- *Nicht-Reflexivität.* Der Akt des Vorhersagens verändert die vorhergesagte Größe nicht. Ändert das Aussprechen der Vorhersage das Verhalten dessen, was vorhergesagt wird, bewegt sich das Ziel, während die Vorhersage gemacht wird.
- *Wiederholbarkeit.* Das Ziel ist eine Instanz einer wiederkehrenden Art, kein singuläres Ereignis, sodass „Wahrscheinlichkeit“ überhaupt einen Referenten hat. Ein Einmaliges hat keine Häufigkeit.

Wo diese gegeben sind, ist probabilistische Vorhersage nicht bloß legitim, sondern mächtig — die versicherungsmathematische Tafel, die gut charakterisierte physikalische Messung, die kalibrierte Wettervorhersage sind hier alle zu Hause. Wo sie versagen, liefert die Methode weiterhin eine Antwort, identisch formatiert, und die Antwort ist unbegründet.

3 Die Regime, in denen die Bedingung versagt

3.1 Nicht-Stationarität

Viele Prozesse, die uns interessieren, ändern ihre eigenen Regeln. Märkte wechseln das Regime; Institutionen organisieren sich um; eine Technologie verändert genau das Verhalten, zu dessen Messung sie gebaut wurde. Ein an das frühere Regime angepasstes Modell beschreibt dann mit ungeminderter formaler Konfidenz eine verschwundene Welt. Die Wirtschaftsgeschichte ist dicht an Beispielen für Modelle, die gut vorhersagten, bis das Regime endete, das sie gültig machte — und die Dichte der vergangenen Daten gibt keine Warnung, weil der Bruch genau das ist, was die Vergangenheit nicht enthält. Die Stationarität ist die am wenigsten geprüfte der Bedingungen, weil ihr Versagen unsichtbar ist, bis es eingetreten ist.

3.2 Reflexivität

Wo das Ziel ein Akteur ist, oder eine Population von Akteuren, die von der Vorhersage erfahren können, stört die Vorhersage ihren Gegenstand. Eine

Prognose über einen Markt kann den Markt bewegen; eine Vorhersage über eine Person, ihr mitgeteilt, verändert, wie sie handelt; eine Projektion der Wirkung einer Politik formt die Politik und damit die Wirkung. Vorhersage und Ziel sind nicht länger unabhängig, und man kann hinterher nie mehr die Treffgenauigkeit der Prognose von ihrem Einfluss auf das Prognostizierte trennen. Dies ist die Domäne, in der Vorhersage nicht bloß unsicher, sondern selbstkorumpierend ist: Sie fabriziert einen Teil ihrer eigenen Bestätigung oder Widerlegung. Reflexivität ist der Grund, warum Vorhersage über menschliche Angelegenheiten kategorisch schwerer ist als Vorhersage über die gleichgültige Natur. Die Natur liest die Prognose nicht.

3.3 Singularität

Manche Fragen sind keine Instanzen einer sich wiederholenden Art. Der Ausgang eines bestimmten beispiellosen Ereignisses, die Trajektorie einer spezifischen neuartigen Situation, das Verhalten eines Systems in einem Regime, das es nie eingenommen hat — diese haben keine Häufigkeit, aus der sich schätzen ließe. Eine Wahrscheinlichkeit, die für einen singulären, beispiellosen Ausgang angeboten wird, ist eine Zahl ohne Nenner: Sie borgt sich die Form häufigkeitsbasierter Inferenz, während ihr die wiederholten Versuche fehlen, die einer Häufigkeit Bedeutung gäben. Die Schätzung kann gleichwohl produziert werden und wird nicht anders aussehen als eine, die auf einer Million Versuchen ruht.

3.4 Das sich auflösende Ziel

Es gibt ein subtileres Versagen, bei dem die vorhergesagte Größe kein Faktum über irgendeinen Fall ist, sondern ein Artefakt des Aggregats — eine zentrale Tendenz, die keine einzelne Instanz verkörpert und die sich in dem Moment auflöst, in dem man auf sie hin handelt, als wäre sie ein Faktum über die vorliegende Instanz. „Den Durchschnittsfall“ vorherzusagen und dann einem besonderen Fall zu begegnen, ist eine alltägliche Weise, auf die eine formal gültige Aggregatvorhersage für genau die Instanz ungültig ist, auf die sie angewandt wird. Der Durchschnitt ist real als Eigenschaft der Population und abwesend als Eigenschaft des Mitglieds.

In allen vier Regimen ist der wesentliche Punkt derselbe, und es lohnt, ihn deutlich auszusprechen: Das Versagen ist in der Ausgabe unsichtbar. Eine Vorhersage aus einem nicht-stationären, reflexiven, singulären oder nur im Aggregat bestehenden Prozess ist genauso formatiert wie eine Vorhersage aus einem stationären, unabhängigen, dicht

beprobten, wiederholbaren. Die Form der Antwort ist konstant über die Grenze ihrer eigenen Gültigkeit hinweg.

4 Warum flüssig formulierende Systeme das Problem verschärfen

Alles Vorstehende gilt für probabilistische Vorhersage im Allgemeinen, bei jedem Schätzer, ob menschlich oder maschinell. Zeitgenössische große Sprachmodelle fügen drei spezifische Verschärfungen hinzu.

4.1 Das konstante Register

Ein menschlicher Prognostiker, der rät, klingt gewöhnlich auch so — voller Vorbehalte, zögernd, sichtbar tastend. Ein statistisches Modell gibt wenigstens eine Varianz aus, ein Konfidenzintervall, einen Fehlerbalken. Ein flüssig formulierendes generatives System gibt weder das hörbare Zögern des Menschen aus noch die explizite Varianz des Statistikers. Seine Ausgabe in einer Domäne, in der Vorhersage gültig ist, ist sprachlich ununterscheidbar von seiner Ausgabe in einer Domäne, in der sie grundlos ist: dieselbe gemessene, kohärente, kompetente Prosa. Die Unsicherheit ist in dem Kanal, den der Nutzer tatsächlich liest, nicht repräsentiert. Dies ist, für das vorliegende Argument, die folgenreichste einzelne Eigenschaft solcher Systeme: Sie können ihre eigene Unsicherheit nicht hörbar machen.

4.2 Interpolation über das Neuartige

Solche Systeme sind darauf trainiert, die zentralen Tendenzen ihres Korpus zu reproduzieren. Mit einem genuin neuartigen Regime konfrontiert — einem, das in den Trainingsdaten fehlt oder dort nur spärlich vertreten ist — meldet das System kein Fehlen; es interpoliert und rendert das Unvertraute in der Gestalt des Vertrauten. Ein Regimebruch, also genau der Fall, in dem die Vergangenheit der falsche Führer ist, ist daher der Fall, für dessen Markierung ein solches System am wenigsten gerüstet ist, denn ihn zu markieren würde verlangen zu erkennen, dass die Gegenwart außerhalb der Verteilung liegt, die das System verinnerlicht hat — und seine Standarddisposition ist es, den eingehenden Fall als aus dieser Verteilung gezogen zu behandeln.

4.3 Die reflexive Schleife im großen Maßstab

Wenn solche Systeme verwendet werden, um das Verhalten genau der Akteure vorherzusagen, die sie konsultieren — Konsumenten, Wähler, Teilnehmer eines Marktes, ein Individuum, das Orientierung über ein anderes Individuum sucht — operiert die reflexive Korruption aus Abschnitt 3.2 in großem Maßstab und mit hoher Geschwindigkeit. Auf die Vorhersage hin wird gehandelt; die Handlung bewegt die vorhergesagte Größe; die Bewegung wird als Daten zurückgelesen; und die Verantwortung für die resultierende Entscheidung diffundiert in das System hinein, das keine trägt. Was eine Wette auf eine stationäre Welt war, wird zu einer Intervention in eine nicht-stationäre, fehletikettiert als deren Beobachtung.

Offen zu haltende Frage: Wenn die Konfidenz eines Systems konstant ist über die Domänen, in denen sie gerechtfertigt ist, und jene, in denen sie es nicht ist, auf wen fällt dann die Last, den Unterschied zu kennen — und lässt sich diese Last von genau den Nutzern abtragen, die nach Situation oder Expertise am wenigsten gerüstet sind, sie zu tragen?

5 Das ehrliche Gegengewicht

Das Argument darf nicht zu viel beweisen. Es wäre falsch zu schließen, dass probabilistische Vorhersage Gewissheit in ihren Daten verlangt oder nur im deterministischen Grenzfall gültig ist. Eine große Menge unsicherer, verrauschter Daten trägt exzellente Vorhersagen: Sterbetafeln, gebaut auf Populationen, von denen kein einzelnes Mitglied individuell vorhersagbar ist; Wettervorhersagen, probabilistisch und unvollkommen, die gleichwohl gut kalibriert und genuin informativ sind; physikalische Messungen mit ehrlichen Fehlerbalken. Was diese teilen, sind nicht gewisse Daten, sondern eine regierte Beziehung zwischen Daten und Ziel — stationär genug, dicht genug beprobt, nicht-reflexiv, wiederholt — zusammen mit der Disziplin der Kalibrierung an den eingetretenen Ergebnissen, die fortlaufend prüft, ob die Schätzungen der Wirklichkeit folgen.

Die Behauptung dieses Papiers ist deshalb nicht, dass Vorhersage Gewissheit braucht. Sie ist enger und, wie wir meinen, genauer: Vorhersage erbt Gültigkeit aus der Struktur des erzeugenden Prozesses; manche unsicheren Domänen besitzen die erforderliche Struktur, und manche scheinen es nur;

und der Anschein ist von innen identisch. Die Aufgabe besteht darin zu wissen, in welcher Domäne man sich befindet — und es ist genau diese Unterscheidung, die ein flüssig formulierendes, selbstgewisses System mit konstantem Register dem Blick entzieht.

Es gibt eine entsprechende Disziplin, die wir benennen, ohne vorzugeben, sie sei leicht zu institutionalisieren: die Konfidenz einer Vorhersage daran zu bemessen, wie gut der Prozess bekannt ist, der sie erzeugt, und nicht an der Flüssigkeit, mit der sie sich ausdrücken lässt; Unsicherheit hörbar zu machen, wo der Kanal sie sonst verbergen würde; und die Grenze zu halten zwischen dem Beschreiben dessen, was in den Daten vorhanden ist, und dem Prognostizieren dessen, was es nicht ist — Ersteres ist der Ort, an dem solche Systeme stark sind, Letzteres der, an dem ihre Stärke zur Gefahr wird.

6 Wo wir damit stehen

Wir haben argumentiert, dass die Gültigkeit einer probabilistischen Vorhersage geborgt ist, nicht intrinsisch; dass sie von einem erzeugenden Prozess nur geerbt wird, wenn dieser Prozess stationär, dicht beprobt, nicht-reflexiv und wiederholbar ist; dass vier breite Regime diese Bedingungen verletzen, während sie die äußere Form der Vorhersage unverändert lassen; und dass flüssig formulierende generative Systeme die Lage verschärfen, indem sie ihr Register über die Grenze hinweg konstant halten, Neuartiges als Vertrautes interpolieren und reflexive Schleifen im großen Maßstab schließen.

Nichts davon ist ein Argument gegen die Verwendung solcher Systeme, wo ihre Bedingungen erfüllt

sind, was oft der Fall und wertvoll ist. Es ist ein Argument gegen einen spezifischen und inzwischen verbreiteten Fehler: die Konstanz der Konfidenz eines Systems mit der Konstanz seiner Gültigkeit zu verwechseln. Die Multiplikation, die Abschnitt 2 eröffnet, ist das Sinnbild der Domäne, in der Vorhersage tragfähig ist — ein bekanntes Gesetz, restlos fortgeschrieben. Die Domänen, in denen am meisten auf dem Spiel steht — der sich verschiebende Markt, die reflexive Population, die singuläre Entscheidung, die besondere Person — sind genau diejenigen, die der Multiplikation am wenigsten gleichen; und ein System, das über sie in derselben Stimme spricht, die es für Arithmetik verwendet, erwirbt dadurch keinen Anspruch auf dasselbe Vertrauen.

Die Unterscheidung zwischen Beschreibung und Vorhersage, zwischen der regierten und der unregierten Beziehung, zwischen der hörbaren und der unhörbaren Unsicherheit ist keine, die gegenwärtige Systeme für uns treffen. Bis sie es tun, muss sie von denen getroffen werden, die sie verwenden — und der erste Schritt ist anzuerkennen, dass es die Grenze gibt und dass nichts an der Oberfläche einer selbstgewissen Antwort verrät, auf welcher ihrer Seiten die Antwort liegt.

Offen zu haltende Frage: Ließe sich ein System bauen, dessen geäußerte Konfidenz damit variierte, wie gut der erzeugende Prozess bekannt ist, statt mit der Flüssigkeit der ausdrückbaren Antwort — und würde ein solches System angenommen, wenn sein hauptsächlich sichtbarer Effekt darin bestünde, genau dort „Ich weiß es nicht“ zu sagen, wo gegenwärtige Systeme am beruhigendsten sind?